中文领域专业术语层次关系构建研究*

朱 惠 杨建林 王 昊
(南京大学信息管理学院 南京 210023)
(江苏省数据工程与知识服务重点实验室 南京 210023)

摘要:【**目的**】对如何从中文非结构化文本获取术语的层次关系进行探讨。【**方法**】从 CNKI 获取数字图书馆学科领域文献，通过术语抽取、术语向量空间模型构建、BIRCH 算法聚类和聚类标签确定构建术语的语义层次结构。【**结果**】构建数字图书馆领域术语的层次结构，并对构建结果进行验证，聚类正确率达到 80.88%，类标签抽取正确率达到 89.71%。【**局限**】对构建效果的验证是通过随机抽样进行的，且仅与一种其他构建方法进行实证比较。【**结论**】应用 BIRCH 算法聚类构建术语层次结构，该方法与 K-means 聚类方法相比具有明显优势，具备较高的执行效率和聚类有效性。

关键词: 术语 层次关系 本体 本体学习 聚类
分类号: TP391

1 引 言

领域术语层次关系是领域知识本体的重要组成部分，它将领域术语分类别按层次进行组织，为领域知识的搜索、重用及进一步理解提供条件。甚至有研究认为本体就是具有包含关系的概念之间的一种层次结构[1-2]。人工构建术语层次结构耗时耗力，且受到领域专家背景知识的限制，缺少客观性和一致性，因此借助知识自动获取方法和技术构建术语层次结构便成为一个新的研究方向。目前，常用的获取术语层次关系的方法之一是基于 Harris 假设的方法[3]。该假设的具体内容是：若两个术语的上下文语境相似，则这两个术语也是相似的[4]。已有学者对该假设进行了验证，并证明是有效的[5]。基于 Harris 假设，可以引入聚类方法构建术语层次结构。

本文试图在建立术语向量空间模型的基础上，将 BIRCH 算法和术语共现理论引入到领域本体的术语层次关系构建中，并通过对术语向量空间模型的优化改进聚类结果，由此形成一种从中文非结构化文本构建领域术语层次关系的具体方法。BIRCH 算法是针对大数据的一种聚类方法，已有学者将其应用在文本聚类、大规模网络数据聚类等方面，但还没发现应用在术语层次构建中，因此，本文尝试引入该算法构建术语的层次结构，并与其他聚类方法进行比较分析。

2 相关研究

国内外已有学者对基于非结构化文本如何获取术语层次关系进行了相关研究。

Sun等[6]集成语义分析和数学统计方法，提出一种监督学习方法获得术语以及术语的层次关系。Hu等[7]探讨如何运用机器学习方法(SVMs和CRFs)将网络百科全书中结构化的知识转化成本体形式。Colace等[8]提出一个融合了语义分析、数学统计等方法的本体学习系统。Meijer等[9]利用词性标注器从语料中抽取术语，利用相关过滤方法获得领域相关度较高的术语，并对术语进行词义消歧，最后基于术语共现关系利用归类技术获得术语的层次关系。De Knijff等[10]利用语法分析器从文本语料中抽取术语，采用归类和层次聚类两种方法获得概念的层次关系。Rios-Alvarado等[2]针对

通讯作者：朱惠，ORCID：0000-0002-2357-1506，E-mail：zhuhui@nju.edu.cn。

*本文系江苏省自然科学基金项目"面向专利预警的中文本体学习研究"(项目编号:BK20130587)和中央高校基本科研业务费专项资金项目"我国图书情报学科知识结构及演化动态研究"(项目编号:20620140645)的研究成果之一。

具体领域的文本语料利用聚类分析、语言模式以及上下文信息构建了术语的层次结构。

季培培等[11]采用多重聚类方法获取术语的层次关系。林源等[12]利用基于规则与统计相结合的方法提取领域术语，并插入到由ODP构建的树中得到领域术语的层次关系。彭成等[13]提出利用确定性退火的多重聚类算法获取术语层次关系的流程。谷俊等[14]提出利用蚁群聚类算法对中文术语进行预聚类，再利用K-means聚类算法对预聚类结果进行聚类获得术语的层次关系。韩红旗等[15]提出基于词形规则模板匹配的术语层次关系抽取方法，实现从科技论文文本中抽取类属关系和整体部分关系。涂鼎等[16]使用主题模型对评论集进行描述选出最具代表性的主题词作为候选术语，进而利用WordNet提取术语间语义关联，最终通过多路聚类获得术语层次关系。李树青[17]提出一种利用引文关键词共现技术自动构建图情学科领域术语层次语义关系的方法。

由上述内容可知，国内外学者尝试采用多种知识自动获取方法和技术构建术语层次结构，其中，聚类方法运用较多，主要有K-means聚类、层次聚类、蚁群聚类、基于确定性退火的聚类等，而且通常是多种聚类方法结合或同一聚类方法多重使用才能达到较好的效果。但这些聚类方法存在以下主要缺陷：不适合大型数据的聚类，例如层次聚类，由于占用内存较大导致在大数据上执行效率较低；不能自动确定聚类数目，例如K-means聚类，需要人工指定聚类数目；离群点和噪声数据对聚类结果产生直接影响，这可能导致局部聚类效果较优，但无法得到较为均匀的聚类结果。

本文首先利用BIRCH算法进行预聚类，进而对预聚类结果进行层次聚类，这样能避免上述聚类缺陷。BIRCH算法由Zhang等[18]于1997年提出，采用聚类特征树存储数据，能诊断离群点和噪声数据、有效解决大数据集的聚类问题、利用贝叶斯信息准则以及类合并过程中类间差异性最小值变化的相对指标确定最优的聚类数目。

3　基于BIRCH聚类的术语层次关系获取方法

本节重点探讨基于BIRCH聚类从非结构化文本获取术语层次关系的方法和过程，并分析术语向量空间的变化对术语层次结构的影响，这里的非结构化文本由期刊论文的标题、摘要和关键词构成。

3.1　术语抽取

科研人员是学科领域术语动态变化过程的直接参与者和见证者，他们撰写的科研文献记载了学科的动态发展过程，文献的关键词则是学科研究内容的凝练，因此，可以从科研文献的关键词中抽取领域术语。

但文献作者给出的关键词具有较大的随意性、不一致性以及误差性，因此，有必要对这些候选术语进行统一规范，以符合同一概念的术语唯一化。

领域术语是专业词汇，必须具有一定的领域认可度，因此，本文采用关键词在所有文档中出现的频数N_k作为筛选条件，即若：

$$N_k \geqslant C \quad (1)$$

则认为该关键词被领域普遍认可，可作为该领域的术语，其中C为词频阈值。

3.2　术语向量空间模型构建

以文档为特征项描述术语形成术语向量空间模型，是后续对术语进行聚类的数据基础。以术语集为词典，借助中文分词工具NLPIR获得文档和术语间的语义关联[19]，构建文档术语频数矩阵，再进行TF-IDF特征项权重计算，得到术语文档权重矩阵。

在术语文档向量空间模型中，测度术语间的亲疏程度是依赖术语在文档中的共现。在较短的非结构化文档中，由于术语量较少，导致术语的共现关系较少，术语文档矩阵较稀疏。而从较稀疏的矩阵中挖掘术语的层次关系，效果可能不尽理想。那么，如何增加术语的共现关系，以使得相应矩阵中的数据更稠密呢？

在术语文档向量空间模型中，文档是术语共现的中介，若术语T1、T2均与文档Di关联，则术语T1与术语T2共现。而文档是由许多词汇构成的，因此，也可认为T1与文档Di的所有wi个词汇产生关联，由此，原来的一个术语文档关联扩展成wi个术语词汇关联。同样，T2也与文档Di的所有wi个词汇产生关联，则T1与T2以词汇为中介产生了共现关系。中介转变后，术语的共现关系将会发生明显的变化：原本具备共现关系的术语，它们的共现关系将保持且共现频数会增加；原本不具备共现关系的术语，若各自关联的文档拥有相同的词汇，则会产生关联，从而具备共现关系[20]。

利用NLPIR以术语集为用户词典对非结构化文

档进行分词，选取其中的名词词汇，并去除停用词和低频词，得到所需词汇。术语通过与其关联的文档找到与其关联的词汇，获得术语词汇关联，由此产生<术语，词汇，共现频数>三元组关系，进一步，笔者引入 Ochiia 系数度量术语与词汇之间关联关系的强弱，形成<术语，词汇，关联系数>三元组关系，构建术语词汇权重矩阵。

3.3　两步聚类

基于术语向量空间模型，采用两步聚类法进行聚类：利用 BIRCH 算法进行预聚类，获得较为"粗糙"的聚类结果，在此基础上利用层次聚类获得术语的层次结构。在聚类过程中，对于不满足聚类结束条件的类别均要再次进行两步聚类，因此整个聚类过程是一个多重两步聚类，如图 1 所示：

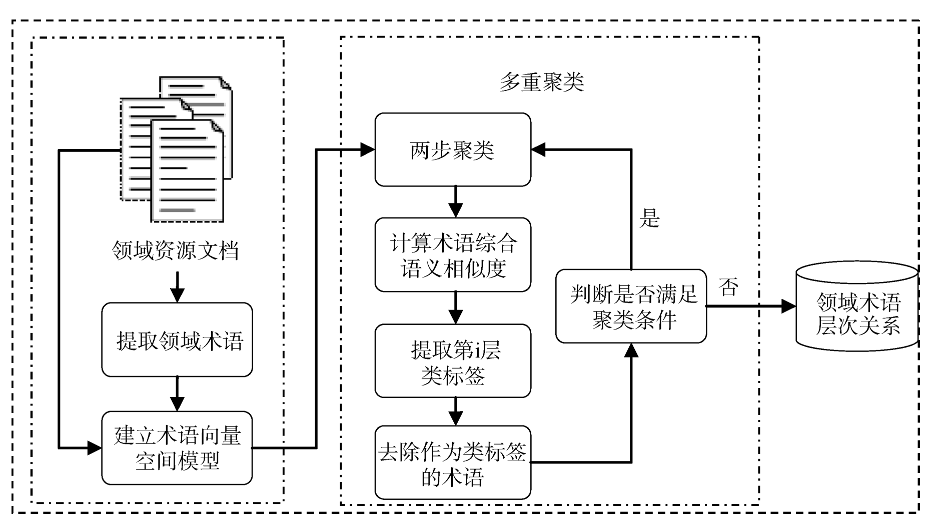


图 1　领域术语层次关系获取方法及流程

BIRCH 算法涉及到两个主要概念：聚类特征 CF(Clustering Feature)和聚类特征树 CF tree。CF tree 中的节点 j 就是类 j，记为 CF_j，包含三个部分：$CF_j=\{N_j,S_{Aj},S_{Aj}^2\}$，其中 N_j 为节点所包含的术语个数，S_{Aj} 为 N_j 个术语的线性和，S_{Aj}^2 为 N_j 个术语的平方和。

例如，假设节点 CF_1 中有三个数据：(1,2)、(3,4)、(5,6),则 $CF_1=\{3,\ (1+3+5,\ 2+4+6),\ (1^2+3^2+5^2,\ 2^2+4^2+6^2)\}=\{3,(9,12),(35,56)\}$。

对于由第 j 类和第 s 类合并形成的新的<j, s>类：

$$CF_{<j,s>}=\{N_j+N_s,S_{Aj}+S_{As},S_{Aj}^2+S_{As}^2\} \tag{2}$$

BIRCH 算法的具体过程如下：

①视所有术语为一个大类，计算 CF，创建根节点；
②读入一个术语，从根节点开始，计算该术语与中间节点(子类)的对数似然距离，并沿着对数似然距离最小的中间节点依次向下选择路径直到叶节点；
③计算术语与子树中所有叶节点的距离，判断最小距离是否小于阈值
　是，则术语被吸收，判断新插入术语的叶节点是否包含足够多的术语
　　是，则分裂该节点，该节点变成中间节点，重新计算叶节点的 CF
　　否，则不分裂该节点
　否，则开辟新的叶节点，重新计算叶节点和所有父节点的 CF
④判断叶节点的数目是否达到最大聚类数目
　是，判断术语是否全部被处理
　　是，结束聚类
　　否，适当增加聚类阈值重新构建较小的 CF tree
　否，判断术语是否全部被处理
　　是，结束聚类
　　否，继续处理下一个术语

两步聚类法在第二步层次聚类过程中通过两个阶段自动确定聚类数目。

(1) 第一阶段，以贝叶斯信息准则(Bayesian Information Criterion, BIC)作为判定标准。

假设聚类数目为 J，则有：

$$BIC(J)=-2\sum_{j=1}^{J}\xi_j+m_J\log N \tag{3}$$

$$m_J=J(2K^A+\sum_{k=1}^{K_B}(L_k-1)) \tag{4}$$

贝叶斯信息准则的第一项即公式(3)反映的是 J 类

对数似然总和，是类内差异性的总度量，第二项即公式(4)是一个模型复杂度的惩罚项，当数据确定后，J越大该项值越大。

若所有样本数据合并成一个大类，此时公式(3)值最大，公式(4)值最小。当预聚类数目增加时，公式(3)值减少，公式(4)值增大，通常增大幅度小于减少幅度，因此总值减少；当预聚类数目增加到 J 时，公式(3)值增大幅度开始大于减少幅度，总值开始增大，此时的J为聚类数目的"粗略"估计值。

(2) 第二阶段，对第一阶段的"粗略"估计值J作修正。用到的指标是：

$$R_2(J)=\frac{d_{min}(C_J)}{d_{min}(C_{J+1})} \quad (5)$$

其中，$d_{min}(C_J)$ 为聚类数目为 J 时，两两类间对数似然距离的最小值。$R_2(J)$ 反映层次聚类的类合并过程中，类间差异性最小值的变化，值越大表明 J+1 类合并到 J 类越不恰当。可依次计算 $R_2(J-1)$、$R_2(J-2)$ 到 $R_2(2)$ 的值，找到其中的最大值和次大值，如果最大值是次大值的 1.15 倍以上，则最大值所对应的J为最终聚类数目，否则，最终聚类数目J为最大值对应的聚类数目和次大值对应聚类数目中的较大值。

3.4 类标签的确定

领域术语层次结构的建立过程也伴随着类标签的确定。本文针对术语层次关系中各层次的各类别，计算类中各术语的综合语义相似度，把拥有最大综合语义相似度的术语提取出来作为类标签[11]。

假设术语 $T_i=(w_{i1}, w_{i2}, \cdots, w_{im})$，术语 $T_j=(w_{j1}, w_{j2}, \cdots, w_{jm})$，则 T_i 与 T_j 的语义相似度定义为：

$$Sim(T_i,T_j)=\frac{\sum_{k=1}^{m} w_{ik}\cdot w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2\cdot\sum_{k=1}^{m} w_{jk}^2}} \quad (6)$$

术语的综合语义相似度是指该术语与类中其他所有术语语义相似度之和。假设类中包含术语 $T_1, T_2, \cdots, T_i, \cdots, T_n$，则术语 T_i 的综合语义相似度为：

$$SumSim(T_i)=\sum_{j=1,j\neq i}^{n} Sim(T_i,T_j) \quad (7)$$

若术语具有最大综合语义相似度，可认为该术语在当前类中代表了最宽泛的语义内容，能作为该类的标签。

4 实验结果及分析

本文以数字图书馆学科领域的期刊论文作为分析对象，基于术语词汇语义关联进行聚类，并对构建的术语层次关系进行有效性验证。

4.1 数据预处理

以"数字图书馆"为主题词，在 CNKI 中国期刊全文数据库的核心期刊范围内检索 1996 年–2011 年期间发表的论文，共计 7 746 篇，抽取标题、摘要和关键词构成非结构化文档。通过术语抽取最终获得 911 个术语，以这些术语为用户词典进行 NLPIR 分词，共得到 50 992 个术语文档关联。若以词汇作为术语的共现中介，通过分词和过滤共获得 2 168 个词汇和 105 477 个术语词汇语义关联。从数据上可以发现术语词汇语义关联数明显大于术语文档语义关联数，语义关联增强，所构建的向量空间也更稠密。

4.2 聚类数目的确定

本文采用的两步聚类法可自动确定聚类数目。设定如下方案：领域专家确定各层聚类数目的取值范围，再由两步聚类法在此范围内自动选出最佳的聚类数目。

假设：n 表示类中的术语数；MaxNum 表示不允许聚类的最大术语数，即若类中术语数小于等于该值，则停止聚类，否则继续；Ceil(X)表示大于等于 X 的最小整数。笔者根据领域特点对各层次聚类数目范围的设定如下：第一层次聚类数目范围为 10-15；第二层次聚类数目范围为 5-10；其后各层次的聚类数目范围与类中包含的术语数目有关：若术语数目大于等于 5×MaxNum，则聚类数目范围为 5-Ceil(n/MaxNum)，否则为 Ceil(n/MaxNum)-5。

4.3 聚类结果分析

针对某个领域，并不知道 MaxNum 取值多少为最佳，因此笔者对 MaxNum 的取值进行多次尝试。令 MaxNum={5,10,15,20}，共进行 4 次尝试，实验结果如表 1 所示。

一个好的聚类层次结构中，整体的深度、宽度以及类内节点数的多少都需较为合理。笔者根据学科领域特点及对聚类结果的观察，最终选定 MaxNum=10。

聚类结果中第 1 层次各类别的相关数据如表 2 所示。

表 1　不同 MaxNum 取值下的聚类结果

指标	MaxNum =5	MaxNum =10	MaxNum =15	MaxNum =20
聚类形成的总簇数	301	190	143	129
第 1 层簇数	10	10	10	10
第 2 层簇数	51	51	51	51
第 3 层簇数	118	96	82	68
第 4 层簇数	96	33	0	0
第 5 层簇数	26	0	0	0
整体最小层次数	4	3	3	3
整体最大层次数	6	5	4	4
类内最多术语数	5	10	15	19
类内最少术语数	1	1	2	3

表 2　聚类第 1 层各类别情况

顶层	第 1 层类号	类标签	术语数	第 2 层聚类数	第 3 层聚类数	第 4 层聚类数	簇数合计
数字图书馆	C1	版权	103	5	10	6	21
	C2	安全	87	5	11	4	20
	C3	知识服务	91	5	10	2	17
	C4	Lib2.0	78	5	7	2	14
	C5	存储	77	5	7	3	15
	C6	个性化	114	5	14	2	21
	C7	评价	126	6	13	8	27
	C8	资源	72	5	8	2	15
	C9	门户	98	5	12	4	21
	C10	多媒体	64	5	4	0	9
	合计	–	910	51	96	33	–

第 1 层类“C3_知识服务”具体层次结构及其包含的部分术语如表 3 所示：

表 3　类“C3_知识服务”的层次结构及其内容

第 2 层	第 3 层	第 4 层	第 2 层	第 3 层	第 4 层	第 5 层
本体			语义网格			
	领域本体			知识管理系统		
	知识共享				数字化权	
	知识库				……	
	知识组织			知识组织系统		
	……				人性化服务	
知识网络					推送技术	
	信息环境				语义互联	
		服务功能			……	
		知识创新		OWL-S		
		知识经济			服务组合	
	运行机制					OWL
		规范控制				本体学习
		知识获取				……
		知识网格		军队院校图书馆		
		……			军队院校	
集成服务					人文关怀	
	网格计算				……	
		信息集成服务	3G			
		移动服务		泛在图书馆		
		可用性			泛在化服务	
		隐私保护			泛在智能	
		……			……	
	信息资源组织			手机图书馆		
		信息服务模式			手机	
		资源配置			无线网络	
		……			……	

聚类分析是一个无监督学习方法，不同参数的设定和实验方案的设计会导致不同的结果。目前还没有统一的标准对聚类结果进行评价，因此，本文通过领域专家对结果进行验证。随机抽取了层次结构中的 10 个父类及其子类，对聚类效果以及类标签抽取的合理性进行考察。

针对抽取出来的每一个父类及其子类：查看子类标签间的关联关系，若大部分的子类标签间具有较强的关联关系，则认为聚类效果较好；查看子类标签与父类标签的关联关系，若大部分的子类标签与父类标签有较强关联关系，则认为类标签的抽取较合理。相关数据如表 4 所示：

表 4　聚类效果及类标签抽取合理性检验

父类编号	父类标签	包含的子类数 S_i	有关联关系的子类数 SSR_i	聚类正确率(%) SSR_i / S_i	与父类标签有关联的子类标签数 SFR_i	类标签抽取正确率(%) SFR_i/S_i
C4	Lib2.0	5	5	100.00	5	100.00
C3	知识服务	5	4	80.00	4	80.00
C4_1	社会阅读	5	4	80.00	4	80.00
C1_4	知识产权	6	4	66.67	4	66.67
C3_1	本体	10	7	70.00	9	90.00
C5_1	网站	9	9	100.00	9	100.00
C8_3_1	数字图书馆建设	7	6	85.71	8	100.00
C9_1_1	资源组织	7	5	71.43	7	100.00
C6_3_4_1	计量分析	7	5	71.43	5	71.43
C7_6_2_2	数字图书馆评价	7	6	85.71	6	85.71
合计	–	68	55	80.88	61	89.71

由表 4 数据可以得出以下结论：

(1) 关于聚类效果。从随机抽取样本的评价结果来看，大部分的类中成员间关系较紧密，聚类正确率均大于等于 66.67%，平均值达到 80.88%。这也反映了本研究所采用的聚类方法能有效针对稀疏数据进行聚类分析。类“C3_知识服务”包含的下层 5 个子类分别为“C3_1_本体”、“C3_2_知识网络”、“C3_3_语义网格”、“C3_5_集成服务”、“C3_4_3G”，易知其中的前 4 个子类间有较强的关联关系，而“C3_4_3G”与其他术语并无明显的关联关系，故排除在外，聚类正确率为 80%。

(2) 关于类标签抽取。在随机抽取的样本中，大部分类的标签抽取较为合理，能与类中较多成员产生关联。类标签抽取正确率均大于等于 66.67%，平均值达到 89.71%。类“C3_知识服务”的 5 个子类中，“C3_1_本体”、“C3_2_知识网络”、“C3_3_语义网格”和“C3_5_集成服务”这 4 个子类的标签与父类标签有较强的关联关系，因此类标签抽取正确率为 80%。

4.4　与 K-means 聚类效果比较

采用 K-means 聚类方法对术语进行层次构建，并与 BIRCH 算法进行比较，具体数据如表 5 所示。

表 5　BIRCH 算法聚类与 K-means 聚类结果比较

指标 \ 聚类方法	BIRCH 算法聚类	K-means 聚类
聚类总簇数	190	398
聚类最深层次数	5	18
类内最少术语数	1	1
类内最多术语数	10	10
平均聚类正确率(%)	80.88	70.39
平均类标签抽取正确率(%)	89.71	55.59

对两种聚类方法的过程和结果进行了比较分析：

(1) BIRCH 算法聚类在确定聚类数目上有优势。K-means 聚类方法需要指定具体的聚类数目，不同的聚类数目确定方案会导致不同的结果，因此，需要花费时间和精力制定合理的方案并进行不断尝试。BIRCH 算法聚类可以在一定的聚类数目范围上根据相关指标自动确定聚类数目。

(2) BIRCH 算法聚类更适合稀疏型数据。K-means 聚类结果中有大量只含有一个术语的类，通过观察，有些术语完全可以并入其他类中，而BIRCH算法聚类的结果中这种现象较少。

(3) BIRCH 算法聚类结果的整体宽度和深度更为合理。K-means 聚类的总簇数达到了 BIRCH 算法聚类的两倍，并且聚类最深层次达到18，这样的聚类结构不能客观合理地反映术语的事实层次关系。

(4) BIRCH 算法聚类的有效性高于 K-means 聚类。通过随机抽取的样本进行计算，K-means 聚类的平均聚类正确率是70.39%，平均类标签抽取正确率是55.59%，低于 BIRCH 算法聚类的80.88%和89.71%。

5　结　语

本文提出一种从领域非结构化文本获取术语层次关系的方法，该方法通过术语抽取、术语向量空间模型构建、BIRCH 算法聚类和聚类标签确定获取术语的语义层次关系。该方法利用术语词汇向量空间代替术语文档向量空间，从而提高了空间的数据稠密度，为后续 BIRCH 聚类的应用提供了良好的数据基础。BIRCH 聚类与其他相关聚类方法相比，具备以下明显优势：适合大数据集的聚类；能诊断离群点和噪声数据；能自动确定聚类数目。本文以数字图书馆领域为例论证了该方法的可行性和有效性，但也存在一些缺陷，对于构建效果的验证只是基于随机抽样进行，且仅与一种其他构建方法进行实证比较。在今后的研究工作中，笔者将进一步尝试运用不同的机器学习方法(半)自动获取领域术语层次关系，探讨更有效可行的策略和方案。

参考文献：

[1] Gruber T R. A Translation Approach to Portable Ontology Specifications [J]. Knowledge Acquisition, 1993, 5(2): 199-220.

[2] Rios-Alvarado A B, Lopez-Arevalo I, Sosa-Sosa V J. Learning Concept Hierarchies from Textual Resources for Ontologies Construction [J]. Expert Systems with Applications, 2013, 40(15): 5907-5915.

[3] 温春，石昭祥，张霄．本体概念层次获取方法综述 [J]．计算机应用与软件，2010，27(9)：103-107. (Wen Chun, Shi Zhaoxiang, Zhang Xiao. A Survey on Ontology Concept Hierarchy Acquisition [J]. Computer Applications and Software, 2010, 27(9): 103-107.)

[4] Harries Z S. Mathematical Structures of Language [M]. New York: Wiley, 1968.

[5] Miller G A, Charles W. Contextual Correlates of Semantic Similarity [J]. Language and Cognitive Processes, 1991, 6(1): 1-28.

[6] Sun C, Zhao M, Long Y J. Learning Concepts and Taxonomic Relations by Metric Learning for Regression [J]. Communications in Statistics-Theory and Methods, 2014, 43(14): 2938-2950.

[7] Hu F H, Shao Z Q, Ruan T. Self-Supervised Chinese Ontology Learning from Online Encyclopedias [J]. The Scientific World Journal, 2014: Article ID 848631.

[8] Colace F, De Santo M, Greco L, et al. Terminological Ontology Learning and Population Using Latent Dirichlet Allocation [J]. Journal of Visual Languages and Computing, 2014, 25(6): 818-826.

[9] Meijer K, Frasincar F, Hogenboom F. A Semantic Approach for Extracting Domain Taxonomies from Text [J]. Decision Support Systems, 2014,62:78-93.

[10] De Knijff J, Frasincar F, Hogenboom F. Domain Taxonomy Learning from Text: The Subsumption Method Versus Hierarchical Clustering[J]. Data & Knowledge Engineering, 2013, 83: 54-69.

[11] 季培培，鄢小燕，岑咏华，等．面向领域中文文本信息处理的术语语义层次获取研究[J]．现代图书情报技术，2010(9): 37-41. (Ji Peipei, Yan Xiaoyan, Cen Yonghua, et al. Research of Term Semantic Hierarchy Induction for Domain-specific Chinese Text Information Processing [J]. New Technology of Library and Information Service, 2010(9): 37-41.)

[12] 林源，陈志泊，孙俏．计算机领域术语的自动获取与层次构建[J]．计算机工程，2011, 37(2): 172-174. (Lin Yuan, Chen Zhibo, Sun Qiao. Computer Domain Term Automatic Extraction and Hierarchical Structure Building [J]. Computer Engineering, 2011, 37(2): 172-174.)

[13] 彭成，季培培．基于确定性退火的中文术语语义层次关联研究 [J]．计算机应用研究，2011，28(9)：3235-3238. (Peng Cheng, Ji Peipei. Research of Term Semantic Hierarchy Correlations Based on Deterministic Annealing [J]. Application Research of Computers, 2011, 28(9): 3235-3238.)

[14] 谷俊，朱紫阳．基于聚类算法的本体层次关系获取研究[J]．现代图书情报技术，2011(12)：46-51. (Gu Jun, Zhu Ziyang. Study on Ontology Hierarchy Relation Induction on Clustering Algorithm [J]. New Technology of Library and Information Service, 2011(12): 46-51.)

[15] 韩红旗，徐硕，桂婕，等．基于词形规则模板的术语层次关系抽取方法[J]．情报学报，2013，32(7)：708-715．(Han

Hongqi, Xu Shuo, Gui Jie, et al. Term Hierarchical Relation Extraction Method Based on Morphology Rule Template [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(7): 708-715.)

[16] 涂鼎，陈岭，陈根才，等. 基于多路层次聚类的商品评论数据概念分类构建[J]. 计算机研究与发展，2013，50(S): 208-215. (Tu Ding, Chen Ling, Chen Gencai, et al. Multi-way Hierarchical Clustering Based Concept Taxonomy Construction for Product Reviews [J]. Journal of Computer Research and Development, 2013, 50(S): 208-215.)

[17] 李树青. 基于引文关键词加权共现技术的图情学科领域本体自动构建方法研究[J]. 情报学报，2012，31(4): 371-380. (Li Shuqing. Research on Automatic Construction of Domain Ontology in Library and Information Science Based on Weighted Co-occurrence of Citation Keywords [J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(4): 371-380.)

[18] Zhang T, Ramakrishnan R, Livny M. BIRCH: A New Data Clustering Algorithm and Its Applications [J]. Data Mining and Knowledge Discovery, 1997, 1(2): 141-182.

[19] NLPIR [EB/OL]. [2014-06-03]. http://ictclas.nlpir.org/docs.

[20] 王昊，苏新宁，朱惠. 中文医学专业术语的层次结构生成研究[J]. 情报学报，2014，33(6): 594-604. (Wang Hao, Su Xinning, Zhu Hui. Study on Hierarchy Structure Generation of Chinese Medical Terminology [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(6): 594-604.)

作者贡献声明：

朱惠：提出研究思路，设计研究方案，进行试验，论文起草及最终版本修订；

杨建林：文献调研，论文修订；

王昊：数据搜集和清洗。

收稿日期: 2015-06-19

收修改稿日期: 2015-09-14

Study on Construction of Domain Terminology Taxonomic Relation

Zhu Hui　Yang Jianlin　Wang Hao

(School of Information Management, Nanjing University, Nanjing 210023, China)

(Jiangsu Key Laboratory of Data Engineering and Knowledge Services, Nanjing 210023, China)

Abstract: [**Objective**] Discuss how to obtain the terminology taxonomic relation from Chinese domain unstructured text. [**Methods**] Based on Digital Library domain text from CNKI, construct terminology hierarchy by terminology extraction, terminology Vector Space Model construction, BIRCH clustering and cluster tag distribution. [**Results**] Obtain the terminology taxonomic relation of Digital Library domain, and evaluate the effectiveness. The accuracy of clustering reaches up to 80.88%, and the accuracy of cluster tag extraction reaches up to 89.71%. [**Limitations**] Evaluate the effectiveness by random sampling, and in comparison with one method only. [**Conclusions**] Making use of BIRCH algorithm to construct terminology taxonomic relation, this algorithm has obvious advantage compared with K-means clustering method, and has higher execution and clustering effectiveness.

Keywords: Terminology　Taxonomic relation　Ontology　Ontology learning　Clustering